IOSR Journal of Mathematics (IOSRJM)
ISSN: 2278-5728 Volume 1, Issue 6 (July-Aug 2012), PP 21-26
www.iosrjournals.org

SomeProperties of SpatialScan StatisticBernoulliModel: ExampleSimulation forSmallandLargeDataUsingSatScan

Titin Siswantining[1*], Asep Saefuddin[2,], A.N.Khairil[2], N. Nunung[2], M. Wayan[2].
[1]*Department of Mathematics , Faculty of Mathematics and Natural Science, University of Indonesia,Depok16424, Indonesia. titin@sci.ui.ac.id*
[2]*Department of Statistics, Faculty of Mathematics and Natural Science, Bogor Agricultural University, Bogor, Indonesia*

Abstract: *Spatial Scan Statisticusing SaTScan is used to detect randomness in the cluster. This paper aims to demonstrate some of the properties owned by the spatial scan statistic of Bernoulli model, which exhibit analytically unbiased, have minimum variance and consistent. Simulations using the SaTScan showed that the amount of data give different results on the cluster properties of randomness.Analysis with small data provides a more random cluster, or unfavorable results of the analysis. It is showed that the greater amount of data provide the LLR enlarge; p-value smaller; RR decreases and Smaller biased.*
Key words: *SpatialScanStatistic, Bernoullimodel,SaTScan, Unbiased, minimumvariance, consistent.*

I. Introduction

Scanstatisticis usedtodetectclustersinthepoint process[1]. SupposeNis aspatialpointprocesswithN(A) is arandomnumber ofpointsinset $A\subset G$.Movingwindowsinthe studyareaisdefinedas a set ofzonesZ, $Z\subset G$.Zare usedinterchangeablytomark thesubsetGandthe set ofparameters defined in thezone. The spatial scan statistic utilise a test statistic based on the ratio maximum likelihood [1,2,3].

SpatialscanstatistichavetwomodelsarethePoissonandBernoulli.ThispaperdiscusesonlyaBernoullimodel.InBernoullimodel, consideringonlythe size ofNsuch thatN(A) is anintegerforallsubsets$A\subset G$.Eachunitsizein accordancewith the entity or individuals whohaveastatementofconditions, e.g.withorwithoutdisease, or includedina particular speciesornot, poor or not. It expressed asindividualdots, andthe locationofindividualsatthepoint. There isexactlyone zone$Z\subset G$such thateachindividualhas achance ofp in thezone, while thechancesof individualsoutside thezoneisq, in the model. Null hypothesisisH_0: p=q, and thealternativehypothesisisH_1: p>q, $Z\in \mathbf{Z}$. Underthe null hypothesisH_0, N(A) ~Bin(M(A), p) forallsetsorthe setA.UnderalternativehypothesisH1, N(A) ~Bin(M(A), p) forallsets $A\subset G$, andN(A) ~Bin(M(A), q) forall subsets$A\subset Z^C$[1].

Somestatisticalmethodsforanalyzingclusterofspatialpointprocessdescribeitmerely,means thatonlycan detectthe location ofthe clusterwithout includinginference, orinferencewithout theability todetect the location ofthe cluster. Importantcharacteristics ofthespatialstatistical testswereboth,such that ifthe null hypothesisis rejectedinthe locations of specific areasthat causerejection.

Asdescribedbefore,itisnecessarytoseehowthestatisticalpropertiesoftheSpatialScanStatisticBernoullimodel analyzed by SaTScan, especially thestatisticderivedfromthe directestimation(DE), whichis usually obtainedfromthesurvey.Statisticalpropertiesbelongingtotheunbiased,minimumvarianceandconsistentlydemonstrated inthispaper. Spatialscanstatistic, especially theconsistentnatureofthe spatialscanstatisticisalwaysobtained when thesize ofa largeexample[1,4], whilein practice due to technical or economic reason, a largesamplesizeis difficult to obtain, so it needsa way outtoovercome this.

1.1. StatisticalProperties of Bernoulli Model

Suppose that therandomvariablehas abinomialdistribution, has aprobabilitydensityfunction (pdf) the following:

$$f(x)=\binom{n}{x}p^x(1-p)^{n-x}, x=0,1,...,n. \tag{1}$$

AccordingtoHoggandCraig[3],unbiasedestimatorwithminimumrangeandaunique(uniqueMVUE=MinimumVarianceUnbiasedEstimate) tobe searchedby using thesufficientstatistic. To findanestimatoris asufficientstatisticornot, can be seenfromthe followingmanner.

Exponentialandlnof(1)is:

$$f(x)=\exp\left[x\ln\frac{p}{1-p}+\ln\binom{n}{x}+n\ln(1-p)\right] \tag{2}$$

Generalformof theexponentialclassis:

$$f(x)=\exp\left[a(p)K(x)+S(x)+b(p)\right]$$

Anexponentialclassof(2), selectedwith $a(p)=\ln\frac{p}{1-p}$, $K(x)=x$, $S(x)=\ln\binom{n}{x}$, and $b(p)=n\ln(1-p)$. According to Hogg and Craig [2005], found that probability density function (pdf) of the exponential class meets

a. the set $\{x \mid x=0,1,...,n\}$ does not depend on p,

b. $a(p)$ function which is continuous and not constant at $0<p<1$,

c. $K(x)$ function is not constant at $x=0,1,...,n$ .probability

Because itsatisfythe abovethree properties, thendensityfunction (pdf) satisfy to theregularcaseofthe exponentialclass. According toHoggandCraig[5], therefore pdfsatisfy to theregularof exponentialclass case, so $K(X)=X$ iscomplete and sufficientstatistic for p.

However, please notethat the

$E(X)=np$ is bias.

Withalgebraic manipulation, is selectedsuch that

$E\left(\frac{X}{n}\right)=\frac{1}{n}E(X)=\frac{1}{n}np=p$ .Proved.

It is mean that $\frac{X}{n}$ is an unbiasedestimatorfor p. Unbiased properties $\frac{X}{n}$ was proven.

Because X iscompletesufficientstatistic, as well as $\frac{X}{n}$ is an unbiasedfor p, thentheLehmann-Scheffetheorem[4], ensuredthat $\frac{X}{n}$ is an unbiasedestimatorand variancethe minimumas well asuniqueto p

Lehmann and Scheffe Theorem.

Supposed that X_1, ..., X_n, n is an integer constant, which have a random sample pdf or PMF f(x;θ), θ∈Ω. For example $Y_1 = u_1$ (X_1, ..., X_n) is asufficient statisticforθ, and suppose thata completefamily $\{f_{Y_1}(y_1;\theta);\theta\in\Omega\}$. And suppose that a complete family$Y_1$is an unique MVUE ofθ.

Furthermore, note that

$$\operatorname{var}\left(\frac{X}{n}\right)=\frac{1}{n^2}\operatorname{var}(X)=\frac{p(1-p)}{n} \tag{3}$$

Thus, the Chebyshev’s inequality is obtained that

$$\Pr\left[\left|\frac{X}{n}-p\right|<k\sqrt{\frac{p(1-p)}{n}}\right]\geq 1-\frac{1}{k^2} \tag{4}$$

Chosed that $k=\varepsilon\sqrt{\frac{n}{p(1-p)}}$ for any $\varepsilon>0$, inequality (4) is transformed into

$$\Pr\left(\left|\frac{X}{n}-p\right|<\varepsilon\right)\geq 1-\frac{p(1-p)}{\varepsilon^2 n} \tag{5}$$

Taking the limit in inequality (5),

$$\lim_{n\to\infty}\Pr\left(\left|\frac{X}{n}-p\right|<\varepsilon\right)\geq 1-\lim_{n\to\infty}\frac{p(1-p)}{\varepsilon^2 n}=1 \tag{6}$$

However, because theprobably is not greaterthan one, then

$$\lim_{n\to\infty}\Pr\left(\left|\frac{X}{n}-p\right|<\varepsilon\right)=1. \qquad (7)$$

Foundthat $\frac{X}{n}$ consistentestimatorfor p . Thus, ifXhas abinomialdistribution, so we obtainedconclusion that $\frac{X}{n}$ is anunbiasedestimatorswithminimumvariance, uniqueandconsistent for p .

1.2. Statistical properties of SaTScan

In the case ofscanstatisticBernoullimodel, therandomvariableisn(G)and sample sizeisN(G), sotheconditions are right(p =q), thenitspdfis

$$f\left(n_G\right)=\binom{N(G)}{n_G}p^{n_G}\left(1-p\right)^{N(G)-n_G}.$$

With thepreviousexplanation, andin a similar way, it was found that $\frac{n_G}{N(G)}$ is anunbiasedestimator withminimumvariance, uniqueandconsistent for p . If$n_z$isthe number ofobservationpointsin thezoneZandif H_1 so, witha similarexplanation, itisobtained that $\frac{n_Z}{N(Z)}$ is an unbiasedestimatorwithminimumvariance, uniqueandconsistent for p , and $\frac{n_G-n_Z}{N(G)-N(Z)}$ isan unbiasedestimatorwithminimumvariance, uniqueandconsistent for q .

Noted that if H_0 ($p=q$)true, then $\frac{n_Z}{N(Z)}$ and $\frac{n_G-n_Z}{N(G)-N(Z)}$ would be consistent to the same value, p . Therefore, the triangle inequality,

$$\begin{aligned}\varepsilon &\le \left|\frac{n_Z}{N(Z)}-\frac{n_G-n_Z}{N(G)-N(Z)}\right| \\ &\le \left|\frac{n_Z}{N(Z)}-p+p-\frac{n_G-n_Z}{N(G)-N(Z)}\right| \\ &\le \left|\frac{n_Z}{N(Z)}-p\right|+\left|\frac{n_G-n_Z}{N(G)-N(Z)}-p\right|\end{aligned},$$

So that

$$\begin{aligned}\Pr\left[\left|\frac{n_Z}{N(Z)}-\frac{n_G-n_Z}{N(G)-N(Z)}\right|\ge\varepsilon\right] &\le \Pr\left[\left|\frac{n_Z}{N(Z)}-p\right|+\left|\frac{n_G-n_Z}{N(G)-N(Z)}-p\right|\ge\varepsilon\right] \\ &\le \Pr\left[\left|\frac{n_Z}{N(Z)}-p\right|\ge\frac{\varepsilon}{2}\ or\ \left|\frac{n_G-n_Z}{N(G)-N(Z)}-p\right|\ge\frac{\varepsilon}{2}\right] \\ &\le \Pr\left[\left|\frac{n_Z}{N(Z)}-p\right|\ge\frac{\varepsilon}{2}\right]+\Pr\left[\left|\frac{n_G-n_Z}{N(G)-N(Z)}-p\right|\ge\frac{\varepsilon}{2}\right]\end{aligned}$$

The resultobtained $\lim_{n\to\infty}\Pr\left[\left|\frac{n_Z}{N(Z)}-\frac{n_G-n_Z}{N(G)-N(Z)}\right|\ge\varepsilon\right]$,

$$\le\lim_{n\to\infty}\Pr\left[\left|\frac{n_Z}{N(Z)}-p\right|\ge\frac{\varepsilon}{2}\right]+\lim_{n\to\infty}\Pr\left[\left|\frac{n_G-n_Z}{N(G)-N(Z)}-p\right|\ge\frac{\varepsilon}{2}\right]=0$$

However, because the chances of non-negative value, then

$$\lim_{n\to\infty}\Pr\left[\left|\frac{n_Z}{N(Z)}-\frac{n_G-n_Z}{N(G)-N(Z)}\right|\ge\varepsilon\right]=0\cdot$$

Orsimply, the valuesof $\frac{n_Z}{N(Z)}$ and $\frac{n_G - n_Z}{N(G) - N(Z)}$ much closed to consistencyfor a large samples.

From this description shows that it is true that the scan statistic have three properties: unbiased, minimum variance, and consistent. Spatial scan statistic requires large sample sizes especially for the consistently properties.

II. Simulation Data BySpatial Scan Statistic

To see how the statistical properties owned by the scan statistic, then performed a data simulation scan statistic and case study analysis. Simulations are carried out by Ugarte [6], which is simulated by dividing the area into 5 zones which percentile to the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and percentile to the 90^{th}. However, in this paper, from 35 area or village observed, divided into four zones, namely p = 0.2; 0.4; 0.8 and p = 0.95. At each village the size of the example let us assume that at 16, so that the paper is distributed Binomial many cases by the number of cases yi = n * pi are many cases in the area to-i is the sample size multiplied by the proportion in thei-thvillage. Results of the spatial scan statistic showed that the proportion of the village with other village is very significant, to the second cluster. It can be seen in Table 1.

Table1. Summary analyzed result byScan statistic simulationof 35 villages.

	Cluster 1 (MLC)	**Cluster 2**
O/E Ratio	1.44	1.44
Risk Ratio (RR)	1.53	1.48
Log Likelihood Ratio (LLR)	16.561325	7.891521
p-value	0.00000034	0.0087
villages	31, 33, 25, 29	22, 30, 32, 35

The observation to the ratio of the expected value is 1.44, means that this estimator have a small bias values. The ratio of observed to expected value equal to 1, means that the estimator be the cluster size of the risk than the risk out of the cluster. The greater value means greater risk in different clusters with the risk outside the cluster. This showsthat theestimatorisclosed tounbiased properties.

The ratio of observed to expected value of the same magnitude, both in cluster 1 (Most Likely Cluster, MLC) and in cluster 2, with a value of RR in MLC was higher than in cluster 2. Very small value of significance it is meaning between the groups with other groups is significantly different proportions.

In order to view further the statisticproperties, so we simulate usingthelarge amounts of data. Forlargeamounts of data, simulationresultsare based onsyntheticestimator, i.e.estimatorsobtainedfromDE,coupledwith theestimationfrom the datavillage potentials (Podes) as anauxiliaryvariable. Simulations summary of the 247 villages syntheticestimator is presented in Table 2. While thecomparison of small data with large data, is summarized in Table 3.

Table2.Simulations summary ofthe247 villagessyntheticestimator

	Thesyntheticestimator247 villages
LLR	413.540690
RR	2.53
O/Eratio	1.53
p-*value*	< 0.000000000000000010
Cluster 1	223, 224, 222, 213, 221, 214, 187, 220, 212, 168, 169, 186, 225, 208, 167, 166, 211, 215, 183, 165, 216, 182, 163, 184, 209, 176, 188, 164,175, 218, 181, 173, 219, 172, 243, 185, 162, 217, 180, 244, 210, 174,179, 239, 246, 177, 242, 178, 161, 247, 245, 171, 198, 238, 205, 204, 241, 61, 199, 158, 240, 170, 237, 206, 194, 13, 235, 59, 197, 160, 193, 157, 234, 154, 236, 232, 195, 60, 62, 200, 231, 155, 230, 233, 85, 203, 53, 58, 189, 228, 63, 159, 229, 156, 227, 151, 57, 84, 192, 226, 153, 82, 201, 70, 207, 202
Cluster 2	136, 137, 138, 147, 146, 149, 134, 140, 135, 145, 139, 141, 144, 133, 150 (O/E = 1.27; RR = 1.3)

Table 3. The simulation summary scan statistic of small and large data

	Cluster 1				**Cluster 2**			
	LLR	O/Eratio	RR	p-value	LLR	O/Eratio	RR	p-value

Small Data	16.561325	1.44	1.53	0.00000034	7.891521	1.44	1.48	0.0087
Large Data	413.54069	1.53	2.53	$< 10^{-16}$	9.237041	1.27	1.30	0.021

In small data, LLRvalueis smaller than thelargeamountsof data, so thatthep-value becomes small, andRRis large,meaning thatthe moresignificantdifferences betweenclustersin to aclusteroutside.
Withsimulateddata,the spatial positioncannot be known;so as tosimulatethe statistical propertiescannot be determined.

III. Acase Study On Poverty In Jember Region Indonesia

Fordata analysisused a case studyon povertyinJember, Indonesia. Summary of the analysisisshownin Table4.

Table 4. Summary of SaTScan results to DE of 35 villages.

	LLR	**RR**	**O/Eratio**	**p-value**	**Villages**
Cluster 1	23.332	2.062	1.711	0.001	19, 8, 34, 5, 4, 26, 11
Cluster 2	16.266788	4.738	4.376	0.001	28, 20

The result of the analysis for a small amount of data is still rather difficult to interpret. Therefore, an analysis was performed for a large amount of data, by using a synthetic estimator. Synthetic estimator obtained from estimating DE with information obtained by borrowing strength from other areas. The information used in this issue is the information on Village Potential (Podes) derived from the Central Bureau of Statistic (BPS) Indonesia [7,8,9]. The analyzed result by using SaTScan,spatialscan statistic software, for the 247 villages is shown in Table 5.
If the twotablesare compared,it appearsthat the greateramount of datawillprovide:
1. LLRenlarge
2. p-value smaller
3. RRdecreases
4. Smallerbias.

The greater ofLLRvalue gives a smaller p-value.

By usingthe casestudies, the spatial positioncan be knownwith certainty. Therefore,the statistical propertiescan be determined.It appears thatwhenlargenumberof data, then theratio ofO/Ebecomes smaller, closeto value1. This means thatunbiased properties can be demonstrated.

Table5.Summary ofSaTScanresultsforthe247syntheticestimator

	LLR	**RR**	**O/E ratio**	**p-value**	**Regions**
Cluster 1	65.647489	1.45	1.19	<0.000000000000000010	234, 13, 235, 85, 170, 233, 63, 236,62, 237, 238, 84, 239, 171, 241, 174,231, 229, 242, 230, 240,82,61,243,60, 57, 69, 173, 81, 245, 175, 228,244, 172, 176, 77,227, 80, 70,232,226, 83, 68, 246, 75, 177, 164, 181,182, 165, 167, 166, 76, 247, 58, 59,180, 159, 79, 64, 91,168,65,67,169, 160, 51, 92, 74, 186, 66, 221,153, 161, 178, 90, 71, 183, 72, 220, 154, 53, 73, 187, 184, 93, 222, 50, 163, 162, 87, 78, 155, 224, 52, 47, 89, 152, 179, 49, 212, 158, 37, 38,208, 188, 213, 88, 185, 143, 157, 193,223, 209, 86, 189, 211, 46, 150, 156
Cluster 2	17.698694	1.74	1.72	0.0000087	95, 94, 96, 101
Cluster 3	15.319119	1.92	1.91	0.000076	126, 127

IV. Conclusion And Suggestions

Spatial scan statistic of Bernoulli model have some properties areunbiased, minimum variance and consistent. Thesepropertiesarequalities thatshould beownedbyan estimator. Based onthe analyticseenthat thespatial scanstatisticwillhavethesepropertiesonlyifa largenumber of samples, particularly on the consistency properties.

Scanstatisticrequireslarge data, whereasinrealitythe available datais small. Analysiswithsmall dataprovidesa morerandomcluster, orunfavorableresultfor the statistical analysis. Therefore needothermethodsthatcanprovidea solutionthe problem of small data.One proposes is touse themethod ofSmallAreaEstimation(SAE).

SAEhas beenknown toovercome theparameterestimateforsmall data[10]. SAEis necessary tojoininto the scanstatisticin order to obtaina betterestimate, comparedwith theinvestigationwithoutjoinedwithSAE, whichrequirelargesamplesizes, wherelargesamplesizesare verydifficult to find.In theory, SAEhas been able tohandle theproblemof smallsamplesize. According toRao [10], SAEalsohas aminimumvariance. Therefore,it is possibletoreplace therole of DEwithSAE estimators.

Refferences

[1]. Kulldorff, M.A Spatial Scan Statistic.*Commun.Statist.-Theory Meth.26(6)*, 1997, 1481 – 1496.

[2]. Duczmal, L., M. Kulldorff, L. Huang. Evaluation of spatial scan statistics for irregularly shaped disease clusters. To appear in *Journal of Computation and Graphical Statistics*.2006.

[3]. Tango, T. A Spatial Scan Statistic with a Restricted Likelihood Ratio.*Japanese Journal of Biometrics*, Vol. 29, No. 2, 2008, 75 – 95.

[4]. Cucala, L., C. Demattei, P. Lopes, & A. Ribeiro. Spatial scan statistics for case event data based on connected components. *Biometrics*, 2009, 1 – 17.

[5]. Hogg, R.V. & A.T. Craig. *Introduction to Mathematical Statistics*. 5th ed. (Prentice Hall, London, 2005)

[6]. Ugarte, M.D., T. Goicoa, A.F. Militino. Empirical Bayes and Fully Bayes procedures to detect high-risk areas in disease mapping. *Computational Statistics and Data Analysis*, *53*, 2009, 2938 – 2949.

[7]. BadanPusatStatistik (BPS, Central Bureau of Statistic) Indonesia.*Basic Poverty Measurement and Diagnostics Course*.(BPS Pubs,Jakarta, Indonesia, 2002).

[8]. BadanPusatStatistik (BPS, Central Bureau of Statistic) Indonesia.*Data and information poverty. 2008*. (BPS Pubs,Jakarta, Indonesia,2008).

[9]. Badan Pusat Statistik(BPS, Central Bureau of Statistic) Indonesia. *Indonesian Statistic2011*. (BPS Pubs, Jakarta, Indonesia, 2012).

[10]. Rao, J.N.K.*Small Area Estimation*. (Wiley-Interscience, USA, 2003)